文章编号:1006 - 9348(2008)03 - 0256 - 03

后验概率支持向量机在企业信用评级中的应用

李 狮[1],夏鹏[2]
(1. 厦门大学信息科学学院,福建 厦门 361005; 2. 浙江师范大学信息工程学院,浙江 金华 321004)

摘要:在支持向量机(Support Vector Machine)的分类问题中,训练样本的分类信息总是确定的,由此得到的分类指示函数也总是对新样本给出确定的分类信息,但是这种情况对一些不确定性问题并不恰当。利用贝叶斯规则,将样本的后验概率与传统支持向量机结合,得到了基于后验概率的支持向量机。在具体的算法上,引入了一个经验性的方法得到样本的后验概率。以某评级机构提供的企业信用评估数据库为研究对象。

关键词:关键词: 支持向量机;后验概率;贝叶斯;非确定性问题; 企业信用评级
中图分类号: TP391; F830 **文献标识码**: B

Application of Posteriori Probability SVM in Enterprise Credit Assessment Model

LI Chong[1], XIA Peng[2]
(1. Department of Automation, Xiamen University, Xiamen Fujian 361005, China;
2. Information Engineering Institute Zhejiang Normal University Jinhua Zhejiang 321004, China)

ABSTRACT: The classified information of the training sample is always certain in the classification problem of support vector machine. The indicator function obtained always gives a certain classification information to the new sample. But it is not appropriate to some uncertain problems. This paper obtains the SVM based on posteriori probability by utilizing the Bayes rule to combine posteriori probability with SVM. An experiential manner is proposed to estimate the posteriori probability of the training data.

KEYWORDS: Support vector machine; Posteriori probability; Bayes; Uncertain classification problem; Enterprise credit assessment

1 引言

经济社会活动中判断一个企业是否守信用牵涉到多数据指标,资产负债率、流动资产周转次数、销售利润率、存贷比、利息偿还率等等,如何从这些数据中判定企业的信用,进而明确地标记出企业的信用等级,更进一步,通过对这些数据的分析,来判断企业将来的信用状况是一个非常复杂的问题。

由于企业财务状况的好坏直接决定了企业的还贷能力,因此考虑的信用评估模型是建立在银行提供的企业财务数据上的。财务数据可分为两类,一类为与行业规模有关的非比率财务指标,另一类为与行业规模无关的财务比率指标。本文主要考虑纯粹的财务比率指标,因为这些指标脱离了具体的企业属性,在不同的企业间具有可比性。通过分析银行提供的财务数据,发现数据属性繁多,而且众多的财务数据分布复杂,对评估模型形成很强的噪声干扰[1]。

SVM 寻找最优超平面,统计学习理论保证其可以具有良好的泛化性能[2] AC框架下得到期望风险的最小化,是一种强有力的学习方法。给定训练样本集:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_l y_l)$$
$$x_i \in R^n, y_i \in \{-1, 1\} \quad (1)$$

其中,为训练样本数服从独立同分布。然而在实际应用中,由于外点(outlier)的存在以及噪声的污染,每个样本对划分的影响应该是不同的。特别是对于不确定性问题,样本只能以一定概率属于某一类,因此用确定的分类信息来表示并不恰当。由于基于后验概率的分类规则比一般的分类规则,能够给出更多的分类信息。结合贝叶斯决策规则,利用后验概率来表示样本的分类信息。

贝叶斯决策理论必须要求概率分布是已知的,因此不能直接应用贝叶斯分类规则。将 SVM 与贝叶斯理论相结合,既利用后验概率表示样本对分类器贡献的差异,以及样本类别信息的不确定性;又避免了对密度函数的估计。

收稿日期:2007 - 05 - 30 修回日期:2007 - 06 - 08

2 后验概率分类问题的表示

本文用后验概率对样本加权,使样本的类别标签不再是 $+1$或 -1,可以将其称为非确定性分类问题。则二分类问题(1)可以表示为:

$$(x_1, p(\omega_i \mid x_1)), (x_2, p(\omega_i \mid x_2)), \cdots, (x_l, p(\omega_i \mid x_l))$$
$$x_i \in R^n, p(\omega_i \mid x_1) \in [0, 1] \quad (2)$$

其中 l为训练样本数目,样本服从独立同分布。ω_i表示某一类,$p(\omega_i \mid x_i)$表示给定 x_i的条件下属于 ω_i类的概率。令 $y_i = 2p(\omega_i \mid x_i) - 1$,则式(2)可表示为:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_l, y_l) \qquad x_i \in R^n, y_i \in \{-1, 1\} \quad (3)$$

可以看出 y_i与 $p(\omega_i \mid x_i)$在表示样本后验概率的意义上等价,而且(1)式可以看作(3)式的一个特例。现假定假设空间是线性函数集,则若存在一组 (w, b),$w \in R^n$, $b \in R$,使得:

$$\begin{cases} wx_i + b > 0, & \text{若 } y_i > 0 \\ wx_i + b > 0, & \text{若 } y_i < 0 \end{cases} \quad 1 \le i \le l$$

那么(3)就可以表示为在后验概率意义下线性可分。

3 最优超平面和间隔定义

3.1 对线性可分情况

考虑式(3),因为训练集中的样本数目是有限的,所以必存在一个 $\triangle > 0$,使 $y_i(wx_i + b) \ge \triangle$成立,即 $y_i\left(\frac{w}{\triangle}x_i + \frac{b}{\triangle}\right) \ge 1$成立,它等价于存在一 (w', b'),使 $y_i(w'x_i + b') \ge 1$成立。SVM 寻找一个最优超平面分隔式(3)分类问题,寻找最优超平面的过程既是寻找最大间隔的过程,令 $\rho(w, b) = \min_{1 \le i \le l}[y_i(wx_i + b)]$,则 $\rho(w, b)$为分类超平面 $wx + b = 0$的间隔,再令 $\|w\| = 1$,使 $\rho = \max_{\|w\|=1}(\rho(w, b))$,则 ρ为分类超平面 $wx + b = 0$的最大间隔,满足条件的 $wx + b = 0$就是最优超平面。

寻找最大间隔的过程等价于下面的优化问题:

$$\max \rho$$
$$\text{s.t } y_i(wx_i + b) \ge \rho, \ \|w\| = 1$$

上面的优化问题等价于:

$$\min \frac{1}{2}\|w\|^2$$
$$\text{s.t } y_i(wx_i + b) \ge 1 \quad (4)$$

根据优化理论,可得式(4)的对偶优化问题:

$$\min \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i, x_j) - \sum_{i=1}^{l}\alpha_i$$
$$\text{s.t } \sum_{i=1}^{l}\alpha_i y_i = \alpha_i \ge 0, 1 \le i \le l \quad (5)$$

则我们可得到最优超平面所对应的 $w_0 = \sum_{i=1}^{l}\alpha_i^0 y_i x_i$,$\alpha_i^0$,为对偶问题(5)的解。$b_0$可由支持向量满足的等式 $y_i(w_0 x_i + b_0) = 1$得出。故优化超平面为:$\sum_{i=1}^{l}\alpha_i^0 y_i (x_i + x) + b_0 = 0$

3.2 对线性不可分情况

引入误差 $\xi = \max(0, 1 - y_i(wx_i + b))$,$1 \le i \le l$考虑软间隔最优超平面的优化问题:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t } y_i(wx_i + b) \ge 1 - \xi_i, \xi \ge 0, \ 1 \le i \le l \quad (6)$$

根据优化理论,可得式(6)的对偶优化问题:

$$\min \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i, x_j) - \sum_{i=1}^{l}\alpha_i$$
$$\text{s.t } \sum_{i=1}^{l}\alpha_i y_i = 0, 0 \le \alpha_i \le C, \ 1 \le i \le l \quad (7)$$

则我们可得到软间隔最优超平面所对应的 $w_0 = \sum_{i=1}^{l}\alpha_i^0 y_i x_i$,$\alpha_i^0$为对偶问题(7)的解。故软间隔最优超平面为:$\sum_{i=1}^{l}\alpha_i^0 y_i (x_i, x) + b_0 = 0$

3.3 对非线性支持向量机的推广

采用核函数 $K(x_i, y_j)$表示高维空间中的内积,则优化问题(8)变为:

$$\min \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{l}\xi_i$$
$$\text{s.t } y_i\left(\sum_{j=1}^{l}K(x_j, x_i) + b \ge 1 - \xi_i, \xi \ge 0, \ 1 \le i \le l\right.$$

其对偶问题为:

$$\min \frac{1}{2}\sum_{i,j=1}^{l}\alpha_i\alpha_j y_i y_j (x_i, x_j) - \sum_{i=1}^{l}\alpha_i$$
$$\text{s.t } \sum_{i=1}^{l}\alpha_i y_i = 0, 0 \le \alpha_i \le C, \ 1 \le i \le l$$

则我们可得到非线性支持向量机为 $\sum_{i=1}^{l}\alpha_i^0 K(x_i + x) + b_0$,$\alpha_i^0$为对偶问题的解。

4 确定样本点后验概率的经验方法

如果知道一个样本点的类条件概率及类先验概率,则由贝叶斯公式我们可以计算出样本点的后验概率,但在实际情况下,并不知道样本点的类条件概率及类先验概率,在此引入一个经验性的方法,在数据集上估计以上两种概率。

对样本点的类概率估计可采用硬临域的方法[3,4]:做超球 $S(x_i, R) = \{x \mid \|x - x_i\| \le R\}$,$R$为某个常数且 $R \le \max_{i,j}\|x - x_i\|$,设 l_i^j为落入 x_i的超球内且属于类 ω_j $(j = 1, 2)$的样本数,l_i为落入 x_i的超球内的样本总数,则类条件概率 $p(x_i \mid \omega_1) = \frac{l_i^j}{l_i}$,根据贝叶斯公式,可计算后验概率为

$$p = (\omega_j \mid x_i) \frac{p(x_i \mid \omega_i)\, p(\omega_i)}{\sum_{j=1}^{2} p(x_i \mid \omega_j)\, p(\omega_j)}$$

虽然该估计值是经验性的,但支持向量机并不像贝叶斯决策那样完全依赖于概率,估计中的误差是可以容忍的。

5 仿真实验

5.1 实验数据准备

本文的实验数据来源于福建省某商业银行 2003年的客户资料,其中有 1147家轻工业企业的财务数据,每条数据包含定量财务指标及银行给定的信用等级 (其中信用等级为 *AAA*的企业 575家,*AA*的企业 471家,*A*及 *A*以下的企业 97家)。从数据显示的信息可以看出,第三类即属于 *A*等级的企业数量明显少于前两类,因此这个问题可以归为不确定性问题之中。根据专家建议,本文仅选择其中的 24个财务比率作为原始特征集并将其编号。其具体含义如表 1所示。

表 1 指标编号

编号	指标名称	编号	指标名称	编号	指标名称
1	资产负债率	9	净资产收益率	17	速动比率
2	流动比率	10	净资产增长率	18	固定资产 / 总资产
3	流动资产周转次数	11	或有负债率	19	息税前收益 / 总负债
4	销售利润率	12	产品销售率	20	息税前收益 / 营运资本
5	总资产报酬率	13	销售收入归行份额与贷款份额比	21	销售收入 / 总资产
6	利息保障倍数	14	存贷比	22	流动负债 / 净资产
7	营运资本比率	15	利息偿还率	23	存货 / 销售收入
8	经营性现金比率	16	到期信用偿还率	24	净现金流量 / 总负债

5.2 实验及结果分析

为了提高准确率,这里使用遗传算法对特征全集进行特征选择,提出 (1, 2, 3, 4, 5, 6, 8, 10, 12, 13, 18, 19, 23) 12个特征进行实验[6]。将全体数据分成两部分:随机抽取 75% 的数据作为训练数据,剩余的 25% 数据作为测试数据。数据经过预处理后,利用第 4节中提到的硬临域方法估算出样本后验概率。在求样本类条件概率的时候所作超球直径 diameter取 0.5,求出样本后验概率 $p(\omega_i \mid x_1)$,使 $y_i = 2p(\omega_i \mid x_1) - 1$。再采用第 3节提到的非线性可分的基于后验概支持向量机进行训练与测试,在具体实验中采用的是径向基核函数。对数据集采用交叉验证 3次得到的最优参数为: (g = 0.13, C = 10)。最终测试分类的正确率为 78.5%。

另外,标准 SVM对参数 C非常敏感,选择不同的 C, SVM分类器的差别非常大,而后验概率支持向量机就避免了这一点。C在很大一个范围内变化对后验概率支持向量机的影响比较小,当 C取 0.1和取 10时候的结果差别不大。这也是后验概率支持向量机的优点之一。

在对比试验中,用标准支持向量机进行训练与测试,经过交叉验证,选取最好的参数 (g = 0.66, C = 32),准确率为 76.3%。

6 结论

在分类问题中采用基于后验概率的支持向量机,能够得到更多的信息,能够避免类别模糊的样本点对分类器的影响。同时对于非确定性分类问题,也更接近真实情况,有较强的适应性,所以在实际问题中可以获得更好的分类准确性。相对于标准 SVM,后验概率支持向量机能够获得更稳定的性能。

参考文献:

[1] 刘闽,林成德. 基于支持向量机的商业银行信用风险评估模型[J]. 厦门大学学报 (自然科学版), 2005, 44(1): 29 - 32.

[2] V Vapnik. The Nature of Statistical Learning Theory. [M] New York: Springer - Verlag, 1999.

[3] 吴高巍,陶卿,王珏. 基于后验概率的支持向量机 [J]. 计算机研究与发展, 2005, 42(2): 196 - 202.

[4] 阎威武,常俊林,邵惠鹤. 一种贝叶斯证据框架下支持向量机建模方法的研究 [J]. 控制与决策, 2005, 19(5): 525 - 528.

[5] Peter solich Probabilistic methods for Support Vector Machines[C] In: Advances in Neural Information Processing Systems 12, S A. Solla, T. K Leen and K - R. Müller (eds), MIT Press, 2000. 349 - 355.

[6] 凌健,林成德. 拆坟特征选择及其在企业信用评估中的应用 [J]. 福建工程学院学报, 2006, 4(4): 436 - 439.

[作者简介]

李 翀 (1983 -),女 (汉族),内蒙古包头市人,硕士生,主要研究方向为数据挖掘与模式识别;

夏 鹏 (1978 -),男 (汉族),天津人,硕士生,主要研究方向为模式识别。